Sussex-Huawei Locomotion Dataset

Data organisation and file formats

D. Roggen, H. Gjoreski, M. Ciliberto, L. Wang

S. Mekki*, S. Valentin*

Wearable Technologies Lab, Sensor Technology Research Centre, University of Sussex

* Mathematical and Algorithmic Sciences Lab, PRC, Huawei Technologies, France

Abstract
This document describes the data organisation and the file formats of the SHL dataset.

1. Revision history

- 03.12.2017: Initial release

2. Data organisation

This applies to all releases of the dataset.

The dataset is organised in a hierarchical directory structure. The data of each user lies in a specific directory for each user (i.e. `User1, User2, User3`).

The recordings of each users are organised in recording sessions with all the files pertaining to that session stored in a sub-directory `ddmmyy` with `dd` the day, `mm` the month and `yy` the year of the recording session. E.g. the directory `120617` corresponds to the recording session on the 12th of June 2017. There may be two recording sessions on the same day, in which case one of the directory is prefixed by an m: e.g. `m280417` for the morning of the 28th of April 2017, and `280417` for the second session recorded in the afternoon.

The data is organised in the following directory structure and with the following files:

/root/	Root folder of the dataset
/root/<userid>/	Root of the User <userid> (i.e. User1, User2, User3).
/root/<userid>/qc.pdf	Quality check document
/root/<userid>/datasetstatus.mat	Convenience Matlab MAT-file containing basic info about the recordings in this session including start time and duration of each session and the quality statistics used to create qc.pdf.
/root/<userid>/labelstatus.mat	Convenience Matlab MAT-file containing basic info about the labels of the recording. This infomation is as well available in the files `labels_track_xxx.txt`.
/root/<userid>/<recordid>/	Root of the record. `<recordid>` is the date of the recording (e.g. ‘120607’). If multiple recordings were done the same day, the one performed in the morning is prefixed

	with m. All the data files recorded on the date <recordid> are under that folder.
/root/<userid>/<recordid>/00inf.txt	Metadata about the recording.
/root/<userid>/<recordid>/<position>_Motion.txt	nsx23 matrix comprising motion and pressure data, with ns the number of samples. `<position>` is one of "Hand", "Bag", "Hips", "Torso". Note that "Hips" refers to the phone work in the trousers front pocket; Torso refers to the phone worn on a breast pocket or on a chest strap (e.g. when running).
/root/<userid>/<recordid>/<position>_API.txt	nsx11 matrix with Google recognition API data
/root/<userid>/<recordid>/<position>_Ambient.txt	nsx6 matrix with Ambient light data
/root/<userid>/<recordid>/<position>_Battery.txt	nsx5 matrix with the Battery data
/root/<userid>/<recordid>/<position>_GPS.txt	nsxvar matrix with the GPS data. The number of columns varies depending on the number of visible satellites.
/root/<userid>/<recordid>/<position>_Wifi.txt	nsxvar matrix with Wifi data. The number of columns is function of the number of nearby Wifis.
/root/<userid>/<recordid>/<position>_Cells.txt	nsxvar matrix with cell data. The number of columns is function of the number of nearby cells.
/root/<userid>/<recordid>/<position>_DeprCells.txt	nsx9 matrix with cell data from the deprecated Android API.
/root/<userid>/<recordid>/<position>_Location.txt	nsx7 matrix indicating latitude/longitude
/root/<userid>/<recordid>/Label.txt	nsx8 matrix with 7 label tracks (Coarse label, Fine label, Road label, Social label, Tunnels label, Traffic label, Food label). The number of samples ns is identical to that in `<position>_Motion.txt` (each line of `<position>_Motion.txt` corresponds to the same line in `Label.txt`).
/root/<userid>/<recordid>/labels_track_main.txt	Contains the fine label track
/root/<userid>/<recordid>/labels_track_road.txt	Road label track
/root/<userid>/<recordid>/labels_track_social.txt	Social label track
/root/<userid>/<recordid>/labels_track_tunnels.txt	Tunnels label track
/root/<userid>/<recordid>/labels_track_traffic.txt	Traffic label track
/root/<userid>/<recordid>/labels_track_food.txt	Food label track
/root/<userid>/<recordid>/GPS.kml	Google Earth GPS trace
/root/<userid>/<recordid>/timelapse.avi	Body-camera timelapse video.

/root/<userid>/<recordid>/videooffset.txt	Offsets allowing to align video time and sensor timestamps.
/root/<userid>/<recordid>/ videospeedup.txt	Constant indicating the timelapse video speedup compared to realtime.
/root/<userid>/<recordid>/data.mat	Convenience Matlab MAT-file containing a data structure with the pre-parsed data of the variable-width files and the data text files of small size.
/root/<userid>/<recordid>/datainf.mat	Convenience Matlab MAT-file containing a data structure with basic info about the recording.

3 File formats

The data of all the motion sensors and the pressure sensors are combined in a single file `<position>_Motion.txt`. This file is resampled on a regular 100Hz sampling grid which is identical for all the `<position>` and for `Label.txt`.

3.1 00inf.txt

Metadata file comprising information about the recording.

Line	What
1	User ID (e.g. User1, User2)
2	timemsmin: first sample time in milliseconds.
3	timemsmax: last sample time in milliseconds.
4	Recording start date in human readable format.
5	Recording length in milliseconds.
6	Recording ID, i.e. name of the folder in which the data is stored (e.g. '120617')

3.2 <position>_Motion.txt

This file contains one line per sample, all sampled at 100 Hz. Some columns may contain NaN if the information is not available (e.g. not all sensors start sampling at the exact same time). The columns are as follows:

Column	What
1	Time ms
2	Acceleration X [m/s^2]
3	Acceleration Y [m/s^2]
4	Acceleration Z [m/s^2]
5	Gyroscope X [rad/s]
6	Gyroscope Y [rad/s]
7	Gyroscope Z [rad/s]
8	Magnetometer X [μT]
9	Magnetometer Y [μT]
10	Magnetometer Z [μT]
11	Orientation w
12	Orientation x
13	Orientation y
14	Orientation z
15	Gravity X [m/s^2]
16	Gravity Y [m/s^2]
17	Gravity Z [m/s^2]
18	Linear acceleration X [m/s^2]
19	Linear acceleration Y [m/s^2]
20	Linear acceleration Z [m/s^2]
21	Pressure [hPa]
22	Altitude derived from the pressure sensor; for all recordings of User1 after 110517 this value is 0.
23	Temperature derived from the pressure sensor; for all recordings of User1 after 110517 this value is 0.

3.3 <position>_API.txt

This file contains one line per sample. The columns are as follows:

Column	What
1	Time ms
2	Ignore
3	Ignore
4	Still confidence (0-100%)
5	On foot confidence (0-100%)
6	Walking confidence (0-100%)
7	Running confidence (0-100%)
8	On bicycle confidence (0-100%)
9	In vehicle confidence (0-100%)
10	Tilting confidence (0-100%)
11	Unknown confidence (0-100%)

3.4 <position>_Ambient.txt

This file contains one line per sample. The columns are as follows:

Column	What
1	Time ms
2	Ignore
3	Ignore
4	Lumix
5	Temperature
6	Ignore

3.5 <position>_Battery.txt

This file contains one line per sample. The columns are as follows:

Column	What
1	Time ms
2	Ignore
3	Ignore
4	Battery level
5	Temperature

3.6 <position>_GPS.txt

This file contains one line per sample. The columns are as follows:

Column	What
1	Time ms
2	Ignore
3	Ignore

4+	Variable number of entries for GPS data. If no satellite is visible the 4th column is 0.
	Otherwise, for each satellite visible 4 columns are added to the data file and an additional last column indicates the number of satellites.
	Each of the 4 columns contain in order: ID, SNR, Azimuth [degrees], Elevation [degrees]
	For example:
	`1489485950011 161777247369 10889909374 0`
	indicates no satellite visible.
	`1489485951014 162780045286 10889909374 7 12.0 56.0 32.0 1`
	indicates one satellite visible; satellite 7 with SNR=12, Azimuth=56 and elevation=32.
	`1489485962025 173791715076 10889909374 7 15.0 56.0 32.0 30 12.0 82.0 70.0 2`
	indicates two satellite visible; satellite 7 and 30.

3.7 <position>_Wifi.txt

This file contains one line per sample. The columns are semicolon delimitated as Wifi SSID can contain spaces. The columns are as follows

Column	What
1	Time ms
2	Ignore
3	Ignore
4+	Variable number of Wifi data. For each visible Wifi 5 semicolon delimited fields are included, in order: BSSID, SSID, RSSI, Frequency [MHz], Capabilities.

3.8 <position>_DeprCells.txt

This file contains one line per sample. Some columns may contain NaN if the information is not available (e.g. the signal is too weak or the cell is not registered). The columns are as follows:

Column	What
1	Time ms
2	Ignore
3	Ignore
4	Network type
5	cid
6	lac
7	dBm
8	MCC
9	MNS

3.9 <position>_Cells.txt

This file contains one line per sample. The columns are as follows

Column	What
1	Time ms
2	Ignore
3	Ignore
4	Number of entries
5+	Variable number of fields depending on entries. The first field identifies the type of cell: LTE, WCDMA or GSM. The number of fields depends on the cell type and is as follows: <LTE; Signal level; Signal strength; Signal level; 28-bit Cell Identity; 3-digit Mobile Country Code; 2 or 3-digit Mobile Network Code; Physical Cell Id; 16-bit Tracking Area Code> <GSM; Signal level calculated based on 3GPP RSRP; Signal strength; Signal level; 16-bit GSM Cell Identity described in TS 27.007; 16-bit Location Area Code; 3-digit Mobile Country Code; 2 or 3-digit Mobile Network Code> <WCDMA; isRegistered; cid; lac; MCC; MNC; PSC; asuLevel; dBm; level>

3.10 <position>_Location.txt

This file contains one line per sample. The columns are as follows

Column	What
1	Time ms
2	Ignore
3	Ignore
4	Accuracy of this location (accuracy as the radius of 68% confidence) [m]
5	Latitude [degrees]
6	Longitude [degrees]
7	Altitude [m]

3.11 Label.txt

This file contains one line per sample. This file is derived from `labels_track_xxx.txt` (with `xxx` being `main`, `road`, `social`, `tunnels`, `traffic`, `food`) and is provided as a convenience. Each line of the file corresponds to the same line in `<position>_Motion.txt`. This may simplify some streaming processing. The column 1 in `Label.txt` is identical to column 1 in `<position>_Motion.txt`.

Column	What
1	Time [ms]
2	Coarse label: Null=0, Still=1, Walking=2, Run=3, Bike=4, Car=5, Bus=6, Train=7, Subway=8
3	Fine label:
	Null= 0
	Still;Stand;Outside= 1
	Still;Stand;Inside= 2
	Still;Sit;Outside= 3
	Still;Sit;Inside= 4
	Walking;Outside= 5
	Walking;Inside= 6
	Run= 7
	Bike= 8
	Car;Driver= 9
	Car;Passenger= 10
	Bus;Stand= 11
	Bus;Sit= 12
	Bus;Up;Stand= 13
	Bus;Up;Sit= 14
	Train;Stand= 15
	Train;Sit= 16
	Subway;Stand= 17
	Subway;Sit= 18
4	Road label: City=1, Motorway=2, Countryside=3, Dirt road=4, Null=0
5	Traffic label: Heavy traffic=1, null=0
6	Tunnels label: Tunnel=1, null=0
7	Social label: Social=1, null=0
8	Food label: Eating=1, Drinking=2, Both=3, null=4

3.12 labels_track_main.txt

This file contains one line per label. The columns are as follows

Column	What
1	Label start time in millisecond
2	Label end time in millisecond
3	Activity label
	Still;Stand;Outside: 0
	Still;Stand;Inside: 1
	Still;Sit;Outside: 2
	Still;Sit;Inside: 3
	Walking;Outside: 4
	Walking;Inside: 5
	Run: 6
	Bike: 7
	Car;Driver: 8
	Car;Passenger: 9
	Bus;Stand: 10
	Bus;Sit: 11
	Bus;Up;Stand: 12
	Bus;Up;Sit: 13
	Train;Stand: 14
	Train;Sit: 15
	Subway;Stand: 16
	Subway;Sit: 17

3.13 labels_track_road.txt

This file contains one line per label. The columns are as follows

Column	What
1	Label start time in millisecond
2	Label end time in millisecond
3	Label: City: 0 Motorway: 1 Countryside: 2 Dirt road: 3

3.14 labels_track_traffic.txt, labels_track_tunnels.txt, labels_track_social.txt

This file contains one line per label. The columns are as follows

Column	What
1	Label start time in millisecond
2	Label end time in millisecond
3	Label: 1=yes (heavy traffic, in tunnel, social interactions) 0=no

3.15 labels_track_food.txt

This file contains one line per label. The columns are as follows

Column	What
1	Label start time in millisecond
2	Label end time in millisecond
3	Label: Eating: 0 Drinking: 1 Both: 2

3.16 timelapse.avi

This is a timelapse video created from images from the body-worn camera.

Two models of body-worn camera were used, one taking a snapshot every 30 seconds, the other taking a snapshot every 32 seconds. A timelapse video has been generated from the camera snapshots, playing at a speed of 0.5 fps (i.e. one frame every 2 seconds). This corresponds to a speedup of 15 or 16 compared to realtime.

3.17 videooffset.txt and videospeedup.txt

These files is used to map video time to sample time.

Definitions (all time units are in milliseconds):

- t_v : video time in [ms]. The first frame of the video has $t_v = 0$
- t_s : sample time in dataset in [ms] from the Unix epoch (Jan 1, 1970). This is the time stored in the first column of the dataset files.
- l : line number (0-based) in the data files containing regular spaced samples at 100Hz. These files are <Bag|Hand|Hips|Torso>_Motion.txt and Label.txt

The file videooffset.txt contains two entries:

$$offset1 \;\; offset2$$

Their meaning is as follows:

- $offset1$: offset between t_s and t_v
- $offset2$: offset between l and t_v

The file videospeedup.txt contains one number: $vspdup$.

$vspdup$ is either 15 or 16 and indicates the speedup of the timelapse compared to realtime:

Video time <> Sample time

The relation between t_v and t_s is: $t_s = offset1 + vspdup \cdot t_v$

Video time <> Line number in Motion/Label files

$$l = offset2 + \frac{vspdup \cdot t_v}{10}$$

3.18 GPS.kml

This is derived data from Hand_Location.txt and labels_track_main.txt. It contains the annotated trajectory of the phone in KML format[1] for visualisation in a software such as Google Earth.

---

1 http://www.opengeospatial.org/standards/kml

3.19 data.mat

This is a convenience Matlab MAT-file which contains the data of the files `<position>_Battery.txt`, `<position>_Ambient.txt`, `<position>_API.txt`, `<position>_Location.txt`, `<position>_Cells.txt`, `<position>_Wifi.txt` in a data structure for easer use.

Loading data.mat yields a data variable which is a structure comprising the fields described in the following table. Each field comprises a 4-entry cell corresponding to each of the 4 body-location of the phones. Locations 1 to 4 correspond to Hand, Bag, Hips and Torso respectively.

Fields	What
Battery{p}	nx5 array comprising the battery data for position p. Format identical to <p>_Battery.txt
Ambient{p}	nx5 array comprising the Ambient data for position p. Format identical to <p>_Ambient.txt
API{p}	nx11 array comprising the recognition API data for position p. Format identical to <p>_Ambient.txt
Location{p}	nx7 array comprising the Location data for position p. Format identical to <p>_Location.txt
Cells{p}	nx7 array comprising the following columns: 1: Time [ms] 2: Ignore 3: Ignore 4 Total number of cells 5 number of LTE cells 6 Number of GSM cells 7 Number of WCDMA
Cellsvar{p}	n-entry cell containing the textual description of the cells. Each entry Cellsvar{p}{i} corresponds to the same entry in Cells{p}{i}. This corresponds to the variable length part of <position>_Cells.txt.
Wifi{p}	nx4 array comprising the following columns: 1 Time [ms] 2 Ignore 3 Ignore 4 Number of Wifi access points
Wifivar{p}	n-entry cell containing the textual description of the Wifi access points. Each entry Wifivar{p}{i} corresponds to the same entry in Wifi{p}{i}. This corresponds to the variable length part of <position>_Wifi.txt.
GPS{p}	nxvar array with GPS data. The columns are as follows: 1 Time [ms] 2 Ignore 3 Ignore 4 Number N of GPS satellites 5-8 Data for 1st GPS: ID, SNR, Azimuth, Elevation 9-12 Data for 2nd GPS, etc.
sessionid	Name of the User (User1, User2 or User3)
timemsmin	Time of start of recording in milliseconds.
timemsmax	Time of end of recording in milliseconds.
recdate	String with date/time of start of recording
reclength	Duration of recording in milliseconds
recid	Recording ID, e.g. '010317'.

3.20 datainf.mat

Convenience Matlab MAT-file containing basic info about the recording. Loading datainf.mat yields a datainfo variable which is a structure comprising the fields described in the following table.

Fields	What
sessionid	Name of the User (User1, User2 or User3)
timemsmin	Time of start of recording in milliseconds.
timemsmax	Time of end of recording in milliseconds.
recdate	String with date/time of start of recording
reclength	Duration of recording in milliseconds
recid	Recording ID, e.g. '010317'.

3.21 datasetstatus.mat

Convenience Matlab MAT-file containing basic info about all the recordings of the user. Loading datasetstatus.mat yields a datasetstatus variable which is a map. The key to the map is the recording ID (e.g. '010317') and the value is a structure comprising the fields described in the following table:

<table>
<tr><td>Fields</td><td>What</td></tr>
<tr><td>sessionid</td><td>Name of the User (User1, User2 or User3)</td></tr>
<tr><td>timemsmin</td><td>Time of start of recording in milliseconds.</td></tr>
<tr><td>timemsmax</td><td>Time of end of recording in milliseconds.</td></tr>
<tr><td>recdate</td><td>String with date/time of start of recording</td></tr>
<tr><td>reclength</td><td>Duration of recording in milliseconds</td></tr>
<tr><td>recid</td><td>Recording ID, e.g. '010317'.</td></tr>
<tr><td>status</td><td>Structure comprising status information as follows:
<table>
<tr><td>Fields</td><td>What</td></tr>
<tr><td>dtok</td><td>Structure indicating if the time interval between samples appears OK according to the automatic checks. This structure contains fields for each sensors (e.g. Accelerometer, Gyroscope). Each field is an 4x1 matrix comprising the status for each of the four phone positions, with index 1 to 4 correspond to Hand, Bag, Hips and Torso respectively.</td></tr>
<tr><td>covok</td><td>Structure indicating if the time sensor coverage appears OK according to the automatic checks.</td></tr>
<tr><td>coverage</td><td>Structure indicating the coverage of the sensor with 1 indicating that the number of samples is as expected and values higher or lower than 1 indicating more or less samples than expected.</td></tr>
<tr><td>boundaryok</td><td>Structure indicating whether the first and last sample of a sensor align with the start and the end of the recording within a tolerance window.</td></tr>
<tr><td>ok</td><td>Structure indicating whether the sensor data is ok; this is a combination of dtok, covok and boundaryok.</td></tr>
<tr><td>restarts</td><td>4x1 matrix indicating the number of restarts of each phones.</td></tr>
<tr><td>motion200hz</td><td>0 if the motion sensors are sampled at 100Hz or 1 if the motion sensors are sampled at 200Hz/250Hz.</td></tr>
</table>
</td></tr>
</table>

3.22 labelstatus.mat

Convenience Matlab MAT-file containing the labels of the recording. Loading labelstatus.mat yields a labelstatus variable which is a map. The key to the map is the recording ID (e.g. '010317') and the value is a structure comprising the fields described in the following table:

Fields	What
main	nx3 array. This is the data of labels_track_main.txt.
food	nx3 array. This is the data of labels_track_food.txt.
road	nx3 array. This is the data of labels_track_road.txt.
tunnels	nx3 array. This is the data of labels_track_tunnels.txt.
social	nx3 array. This is the data of labels_track_social.txt.
traffic	nx3 array. This is the data of labels_track_traffic.txt.
dstill	Duration of still in the recording in milliseconds
dwalk	Duration of walk in the recording in milliseconds
drun	Duration of run in the recording in milliseconds
dbike	Duration of bike in the recording in milliseconds
dcar	Duration of car in the recording in milliseconds
dbus	Duration of bus in the recording in milliseconds
dtrain	Duration of train in the recording in milliseconds
dsubway	Duration of subway in the recording in milliseconds
dtow	Total duration of annotations in the recording in milliseconds.
sessionid	Name of the User (User1, User2 or User3)
recid	Recording ID, e.g. '010317'.